一种新的多标签数据集转换方法 RAPC-W*

兰浩良[1], 朱玉全[1], 陈 耿[2]
(1. 江苏大学 计算机科学与通信工程学院, 江苏 镇江 212013; 2. 南京审计学院 信息科学学院, 南京 211815)

摘 要: 针对现有多标签数据集转换方法无法有效利用标签间的语义相关性和共现性知识,以及转换得到的数据集相对于问题规模偏小等问题,提出了一种新的多标签数据集转换方法 RAPC-W(ranking by all pairwise comparision based WordNet)。该方法将标签对从原来的两对扩展到四对,增加了划分后数据集的规模。另外,引入了外部数据源 WordNet,较好地考虑了标签语义相关性和共现性知识,一定程度上过滤掉了语义不相关的标签组合,更好地保留了原始数据集的信息,降低了噪声数据集对基分类器训练的不良影响。在 UCI 知识库提供的 Yeast 和 Letter 数据集以及 KEEL 提供的 Emotion、Genbase 数据集上的一系列实验结果表明,该方法是有效可行的。

关键词: 多标签; 数据集转换; 相关性; 共现性; WordNet
中图分类号: TP311 **文献标志码**: A **文章编号**: 1001-3695(2013)06-1692-04
doi:10.3969/j.issn.1001-3695.2013.06.023

New multi-label data set division method RAPC-W

LAN Hao-liang[1], ZHU Yu-quan[1], CHEN Geng[2]
(1. *School of Computer Science & Telecommunications Engineering, Jiangsu University, Zhenjiang Jiangsu* 212013, *China*; 2. *School of Information Science, Nanjing Audit University, Nanjing* 211815, *China*)

Abstract: Existing multi-label data set transformation method can not effectively utilize semantic correlation between the label and the co-occurrence of knowledge, as well as the dataset is small-scale to the scale of the problem such, this paper presented a new multi-label data set into method RAPC-W(ranking by all pairwise comparision based WordNet), the method extend the label pair from the original two pairs to four pairs increasing the size of the devided data set. On the other hand introduced external data sources WordNet, taking a comprehensive consideration of the label semantic correlation and the co-occurrence of knowledge, to some extent, filter out the uncorrelated label combination in semantics, better to retain the information of the original data set, also reduce the adverse effects of the noise data set to the based classifier training. A series of experimental results based on the Yeast and Letter data set provided by the UCI knowledge as well as the Emotion and Genbase data set provided by the KELL shows that this method is effective and feasible.

Key words: multi-label; data conversion; correlation; co-occurrence; WordNet

0 引言

随着多标签学习在图像及视频语义标注、功能基因组、音乐情感分类和营销指导等方面的成功应用,多标签学习已经成为数据挖掘领域的一个研究热点。对于多标签学习问题,其处理方法可分为整体优化法和基于数据分解的方法[1]。整体优化法对所有样本和标签构建一个优化问题,如 BoosTexter 算法[2]、Rank2SVM 算法[3]、多标签 K 近邻算法[4]和最大化熵的多标签算法(MIME)[5]等。该方法的优点是没有改变数据的结构,没有破坏类与类之间的联系;其缺点是需要花费大量时间去解优化问题,难以应用到较大规模的数据集。基于数据分解的方法将多标签学习任务转换为一个或多个单标签学习任务,利用已有的单标签数据挖掘知识进行多标签问题的处理。这种转换实际上是先将多标签学习任务中的多标签数据集转换为单标签数据集,再利用 SVM[6]等分类算法在转换后的数据集上进行基分类器[7]的训练,并借助基分类器完成多标签数据的分类工作[8]。要想提升该类方法的准确率,可以在两方面努力:a)寻求有效的基分类器构造方法,在转换后的数据集构建更加高效的基分类器;b)在多标签数据集的转换上下功夫,即可以寻找一种高效的数据集转换方法,使转换后的数据集能更好地反映原始多标签数据集中的信息,从而使得建立在这种数据集上的基分类器具有更高的分类准确率。本文重点从 b)入手,寻找更加有效的数据集转换方法。

目前可用的转换方法主要有 BR 方法[8]、Copy 方法[8]、LP 方法[9]、RPC 方法[10]等。BR(binary relevance)方法是一种典型的基于数据分解的方法,它将每个标签的预测看做一个独立的单分类问题,并为每个标签训练一个独立的分类器,用全部的训练数据对每个分类器进行训练。这种算法忽略了标签之间的相互关系,往往无法达到令人满意的分类效果。文献[8]通过拷贝(copy)和带权重拷贝(copy-weight)的方法对 BR 进行改进,将原训练集合中的一条多标签数据拆分成多条单标签数据,并给予相应的权重;LP(label powerset)是另外一种被广泛

收稿日期: 2012-09-27; **修回日期**: 2012-11-07 **基金项目**: 国家自然科学基金资助项目(71271117);江苏省科技型企业技术创新资金资助项目(BC2012331)

作者简介:兰浩良(1986-),男,山东德州人,硕士研究生,主要研究方向为数据挖掘、模式识别(lanhaoliang200816@126.cm);朱玉全(1965-),男,江苏常州人,教授,博导,博士,主要研究方向为人工智能、数据库系统及其应用等;陈耿(1965-),男,教授,博士,主要研究方向为数据挖掘、审计风险管理等.

使用的转换方法,它将训练数据中的每种标签组合进行二进制编码,从而形成新的标签。LP 算法的显著缺点是不能预测新的标签组合。为此,Read[9]将概率分布模型应用到 LP 中,当对未分类数据进行预测时,可以预测出训练集合中未出现的标签组合。LP 算法的复杂度较高,高达 $O(\min\{2q, m\}\#t(D))$,可以通过剪枝或随机标签组合[11]的方法在一定程度上降低复杂度,但降低的幅度有限;RPC 是 Hullermeier 等人提出的一种方法,该方法是一种基于标签对比(pairwise comparison)的转换方法,通过对比标签集合中任意两个标签之间的关系,建立 $q(q-1)/2$ 个分类器。每个分类器在两个标签 λ_i 和 λ_j 间投票,然后组合这些投票结果作为最终的多标签分类结果[12]。假设多标签分类算法中采用的基础分类器(base classifier)的复杂度为 $O(t(D))$,其中函数 $t(D)$ 表示分类器在训练集合 D 上建立分类模型的复杂度,则基于标签对比的多标签分类算法的复杂度为 $O(q(q-1)/2\#t(D))$。

上述这些方法均没有很好地利用标签之间潜在的语义相关性和共现性知识,并且转换得到的数据集相对于问题本身规模偏小,不能进行有效的基分类器训练。为此,本文在 RPC 方法的基础上,结合外部数据源 WordNet,提出了一种新的多标签数据集转换方法 RAPC-W。

1　相关知识介绍

1.1　外部数据源 WordNet

在进行语义相似性和相关性度量方面,基于结构化语义知识库 WordNet 的度量一直在这一领域独领风骚,形成了一系列较为成熟的方法。它将名词、动词、形容词和副词分别按照词义进行组织,形成同义词集(synsets)。同义词集之间通过多种语义关系进行连接,其中最基本的语义关系是上下位关系。根据上下位关系可以将 WordNet 的同义词集形成一个树状的层次结构,各种语义相关性的计算大多是基于这个层次结构进行的。下面介绍几种基于 WordNet 的词间语义相关性度量方法。

1)基于文本重叠度的度量　Lesk[13]提出使用单词定义的文本重叠度进行相关性的计算。Banerjee 等人对 Lesk 的方法进行了改进,提出了扩展的注释重叠度方法,效果较前者有所改进。

2)基于路径的度量　Wu 等人考虑了两个概念节点及其最小公共父节点在层次结构中的深度来计算语义相似性,具体计算公式如下:

$$\mathrm{Kwnc}(c_1,c_2)_{w\cup p}=\frac{2\times d(\mathrm{lcs})}{l(c_1,\mathrm{lcs})+l(c_2,\mathrm{lcs})+2\times d(\mathrm{lcs})} \tag{1}$$

其中:lcs 为节点 c_1 和 c_2 在 WordNet 层次结构中的最小公共父节点;$d(\mathrm{lcs})$ 为节点 lcs 在层次结构中的深度;$l(c_1,\mathrm{lcs})$ 为节点 c_1 到 lcs 的距离。

Leacock 等人综合考虑了层次结构的深度和节点间的距离对相关性的影响,提出了一种规格化的测量相似性的方法,具体计算公式如下:

$$\mathrm{Kwnc}(c_1,c_2)_{\mathrm{LCH}}=-\log\frac{l(c_1,c_2)}{2D} \tag{2}$$

其中:$l(c_1,c_2)$ 为节点 c_1 与 c_2 之间的距离,D 为 WordNet 层次结构的最大深度。

3)基于信息量的度量　Resnik 根据概念在一个文集中的出现概率度量其信息量,由两个概念的最小公共父节点的信息量计算概念对之间的相关性。两个概念间共享信息的程度为

$$\mathrm{Kwnc}(c_1,c_2)_{\mathrm{RES}}=IC(\mathrm{lcs}) \tag{3}$$

$$IC(\mathrm{lcs})=-\log P(\mathrm{lcs}) \tag{4}$$

其中:$P(\mathrm{lcs})$ 为概念 lcs 出现的概率;$IC(\mathrm{lcs})$ 为概念 lcs 的信息量。

Jiang 等人[14]提出了一种新颖的基于信息量和文集混合测量语义距离的 JNC 度量,该方法被认为是最有效的一种度量方法。JNC 中需要给定一个已经按照语义进行过改良好分类的词库,由这个词库可以估计出每个概念 c 出现的概率。

$$Pr(c)=\frac{\mathrm{Freq}(c)}{N} \tag{5}$$

其中:$\mathrm{Freq}(c)$ 指概念 c 出现的次数,N 是所有概念出现的总次数。利用此概率可定义该概率对应的信息容量为

$$IC(c)=-\log(Pr(c)) \tag{6}$$

基于 WordNet 的语义相关性度量可理解为两个概念分别去除公共概念之后信息容量和的倒数,即

$$\mathrm{Kwnc}(c_1,c_2)\frac{1}{IC(c_1)+IC(c_2)-2IC(\mathrm{lcs}(c_1,c_2))} \tag{7}$$

其中:$\mathrm{lcs}(c_1,c_2)$ 是概念 c_1 与 c_2 的最低公共父节点,可由 WordNet 的概念树结构来得到。

有了词汇间关系最有效的度量,就能利用式(5)~(7)来定量地计算两两标签之间的关系量。

获得了两两标签之间的关系量之后,在进行多标签数据集转换时利用获得的关系量来定量地度量要合并的两个标签之间的关系,若关系量的值大于用户指定的合并阈值 λ,则将两个标签合并;否则舍弃。

1.2　RPC

RPC(ranking by pairwise comparison)将多标签数据集转换成 $M(M-1)/2$ 个二标签数据集,对于每一组标签(λ_i, λ_j),$1\leqslant i<j\leqslant M$,每一个数据集中所包含的是原始数据集中至少包含这两个标签中的一个标签,但不同时包含两个标签的样本(只包含 λ_i 或 λ_j)。然后从这些数据集中学习一个针对这两个标签的二分类器。给一个测试样本,所有的二分类器都会被调用。每个分类器在两个标签 λ_i 和 λ_j 间投票,然后组合这些投票结果作为最终的多标签分类结果。假设多标签分类算法中采用的基础分类器的复杂度为 $O(t(D))$,其中函数 $t(D)$ 表示分类器在训练集合 D 上建立分类模型的复杂度,则基于标签对比的多标签分类算法的复杂度为 $O((q(q-1)/2)\times t(D))$。

当前多标签学习的重点还在于如何提高多标签分类的准确率,RPC 方法虽然在时间复杂度上较 Copy、BR、LP 等方法有所改进,但其在多标签分类上的性能还有待挖掘,该方法没有很好地利用标签之间潜在的语义相关性和共现性知识,并且转换得到的数据集相对于问题本身规模偏小,不能进行有效的基分类器训练。

2　基于 RPC 多标签数据集转换方法 RAPC-W

RAPC-W 方法是在 RPC 方法的基础上结合外部数据源 WordNet 提出的。该方法将数据集 S 中的所有标签(C_1, C_2, C_3, …, C_n)进行两两组合,并且取任意两个标签组合(C_i, C_j)的四种情况:(C_i, C_j)、($\neg C_i$, C_j)、(C_i, $\neg C_j$)、($\neg C_i$, $\neg C_j$),依据这四种情况进行数据集的划分,具体地说就是当数据集中的

数据项出现以下两种情况中的任意一种时才对该数据项进行标记,否则不予标记,即舍弃该数据项。其具体步骤如下:

a)当且仅当一条数据项中只含有标签 C_i 或只含有标签 C_j 时,将该条数据标记为(X,$C_i\neg C_j$)或标记为(X,$\neg C_iC_j$)。

b)当且仅当一条数据项中同时含有 C_iC_j 或同时不含有 C_iC_j(也就是含有$\neg C_i\neg C_j$),并且利用式(5)~(7)计算获得的 Kwnc(C_i,C_j)大于设定的阈值 λ 时,即 Kwnc(C_i,C_j) > λ 时,将这条数据标记为(X,C_iC_j)或标记为(X,$\neg C_i\neg C_j$)。

具体描述如算法 1 所示。

算法 1 基于 RPC 与 WordNet 的多标签数据集转换方法 RAPC-W

输入: 多标签数据集 S;阈值 λ。

输出: 单标签数据集 sim_data。

```
1 初始化新的数据集 sim_data;
2 for 1 to M //M 为标签个数
3 for j = (i +1) to M
4 P(CiCj) ← getPairLabel(S);//获取所有的标签对组合
5     end for;
6 end for;//转换 S 中仅含单标签的数据集
4 for each CiCj in P(CiCj) do
5 data (Ci) ← F(S,Ci);//获得含有 Ci 不含 Cj 的数据项
6 (X,Ci¬ Cj) ← M(data(Ci));//标记含有 Ci 的数据项
7 data (Cj) ← F(S,Cj);//获得含有 Cj 不含 Ci 的数据项
8 (X,¬ CiCj) ← M(data(Cj));//标记含有 Cj 的数据项
9 初始化一个单标签数据集子集 sim_dataSet;
10 sim_data ← (X,Ci¬ Cj);//将新数据项放入 sim_data 中
11 sim_data ← (X,¬ CiCj);//将新数据项放入 sim_data 中
12 sim_data ← sim_dataSet;//将数据子集放入 sim_data 中
13 end for;//转换 S 中含一对标签或一对均不含的数据集
14 for each CiCj in P(CiCj) do
15 计算 Knwc(Ci,Cj);//通过式(5)~(7)
16 if Knwc(Ci,Cj) >λ then
17 data (CiCj) ← F(S,CiCj);//获得含 CiCj 的数据项
18  (X,CiCj) ← M(data (CiCj));//标记含 CiCj 的数据项
19 data(¬ Ci¬ Cj)← F(S,¬ Ci¬ Cj);/ * 获得不含 CiCj 的数据项 * /
20 X,¬ Ci¬ Cj) ← M(data (¬ Ci¬ Cj));/ * 标记不含 CiCj 的数据项 * /
21 初始化一个单标签数据集子集 dou_data;
22 dou_data ← (X,CiCj);//将数据项放入 dou_data 中
23 dou_data ← (X,¬ Ci¬ Cj);//将数据项放入 dou_data 中
24 sim_data ← dou_data;//将数据子集放入 sim_data 中
25 end for;
```

3 实验结果与分析

本文对 RPC 与 PAPC-W 进行实验对比时,借助两种方法在转换后的数据集上进行基分类器的训练,第一种方法采用的 SVM 训练程序来自台湾大学林智仁教授等人开发的 Libsvm,SVM(support vector machine)分类器采用线性核函数,该方法记为 SVM_L;第二种方法采用的是多项式核函数,常数 c 的值设为 1,该方法记为 SVM_D。采用的数据集为 UCI 知识库提供的 Yeast 和 Letter 数据集以及 KEEL 提供的 Emotion、Genbase 数据集,基本信息如表 1 所示。其中,平均标签长度为数据集中每个样本所拥有的标签个数的平均值。

表 1 实验数据集信息

数据集	样本个数	样本特征个数	标签数	平均标签长度
Yest	2 417	103	14	4.25
Emotion	593	72	6	1.87
Genbase	662	1 185	27	1.35
Letter	20 000	16	26	1

本文从准确率、灵敏性、特效性、精度、F1 值、NumberWords 六个方面对所提出的 PAPC-W 方法与传统 RPC 方法进行比较。灵敏性也称真正(识别)率,即正确识别正元组的百分比;特效性是指真负率,即正确识别负元组的百分比;准确率是正确识别正负元组的百分比,它是灵敏性和特效性的函数;精度是指正确识别的正元组在识别为正元组中所占的比率;F1 值是精度和灵敏性的调和平均数;NumberWords 即至少被正确识别一次的标签数量,这一数值反映了转换方法对标签的覆盖程度,记为 NumWords。

采用 SVM_L 作为基分类器,在数据集 Genbase、Letter、Emotion和 Yeast 上的实验结果分别如图 1~4 所示;采用 SVM_D 作为基分类器,在数据集 Genbase、Letter、Emotion 和 Yeast 上的实验结果分别如图 5~8 所示。其中,+、- 和 ~ 分别表示本文方法优于、劣于和持平于对比方法。

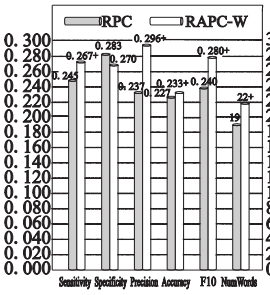


图1 Genbase上的对比结果

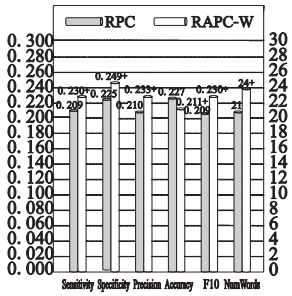


图2 Letter上的对比结果

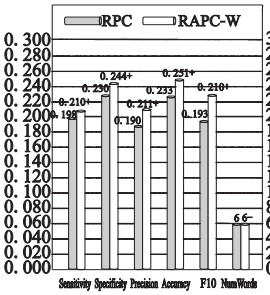


图3 Emotion上的对比结果

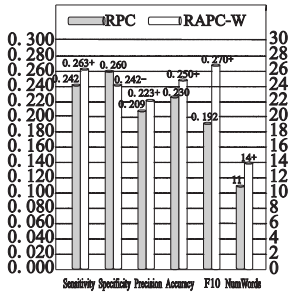


图4 Yeast上的对比结果

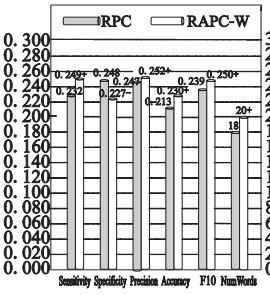


图5 Genbase上的对比结果

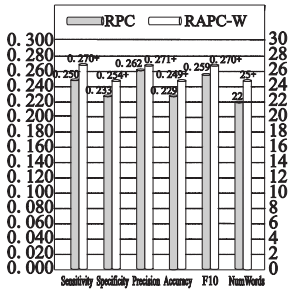


图6 Letter上的对比结果

借助两种基分类器,在四个多标签公用数据集上对 RAPC-W 方法与 RPC 方法进行比较,对比结果显示:图 1、4、5 中的 Specify 指标以及图 2 中的 Accuracy 指标,RAPC-W 方法较 RPC 方法有不同程度的下降;图 4 与图 8 中的 NumberWords 指标,RAPC-W 方法与 RPC 方法持平。除此之外,各数据集上的各项指标,RAPC-W 方法较之 RPC 方法均有明显提高,至少在 80% 的性能指标上,RAPC-W 较 RPC 有显著提高。

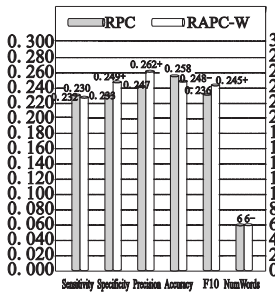


图7　Emotion上的对比结果

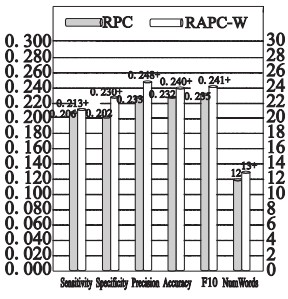


图8　Yeast上的对比结果

4　结束语

本文在多标签数据集划分方法 RPC 的基础上,提出了一种新方法 RAPC-W,在标签组合时引入了 WordNet 校验,通过在 UCI 知识库以及 KEEL 提供的数据集上的一系列实验,验证了本文所提方法的有效性。需要指出的是,本文在对标签组合引入 WordNet 校验时,只是进行了简单的不相关词过滤,并未有效利用数据集中的上下文信息。因此,如何利用数据集中的上下文信息并结合 WordNet 进行标签组合的过滤,是下一步值得深入研究的工作。

参考文献:

[1] SOUMAKAS G T. Multilabel classification: an overview[J]. International Journal of Data Warehousing and Mining,2007,3(3):1-13.

[2] SANTOS A M. Analyzing class ification methods in multi-label tasks [M]. Berlin:Spring-Verlag,2010:137-142.

[3] SCHAPIRE R E. Boostexter: a boosting based system for text categorization[J]. Machine Learning,2000,39(2-3):135-168.

[4] ZHANG Min-ling,ZHOU Zhi-hua. A K-nearest neighbor based algorithm for multi-label classicication[C]//Proc of IEEE International Conference on Granular Computing. 2004:718-721.

[5] ZHU Sheng-huo,JI Xiang,XU Wei,*et al*. Multi-labelled classification using maximum entropy method[C]//Proc of the 28th Annual International ACM SIGIR Congerence on Reserch and Development in Information Retrieval. New York: ACM Press,2005:274-281.

[6] CHANG C C. A library for support vector machine[EB/OL]. (2010-01-10). http://www. csie. ntu. twl/ ~ cjlin/libsvm.

[7] TSOUMAKAS G, DIMOU A, SPYROMITROS E, *et al*. Correlation-based pruning of stacked binary relevance models for muti-label learning[J]. Machine Learning,2009,54(1):5-32.

[8] TROHIDIS K. Multi-label classification of stack binary relevance models for mutilabel classifiers [J]. Eurasip Journal on Audio Speech and Music Processing,2011,4(1):4-6.

[9] READ J. A pruned problem transformation method for multi-label classification[C]//Proc of New Zealand Computer Science Reasearch Student Conference. 2008:143-150.

[10] TSOUMAKAS G, KATAKIS L, VLAHAVAS L. Mining multi-label data[C]//Proc of Data Mining and Knowledge Discovery Handbook. 2010:667-685.

[11] SECHIDIS K,TSOUMAKAS G, VLAHAVAS I. On the stratification of multi-label data[C]//Proc of European Conference on Machine Learning and Knowledge Discovery in Databases. 2011:145-158.

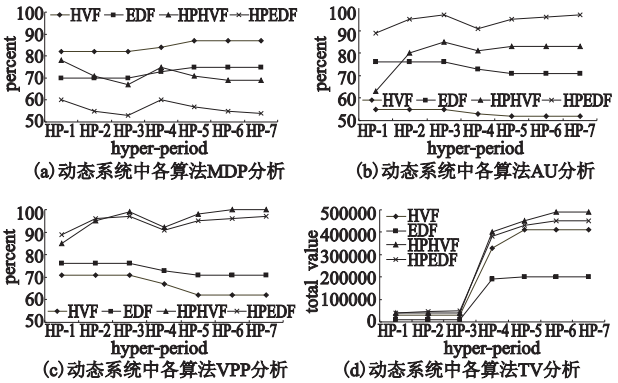
[12] READ J. Classifier chains for multi-label classification[J]. Machine Learning Journal Springer,2011,85(3):333-359.

[13] LESK M. Automatic sense disambiguation using machine readable dictionaries: hoto tell a pine cone from an ice cream cone[C]//Proc of the 5th Annual International Conference on Systems Documentation. New York: ACM Press,1986:24-26.

[14] JIANG J J,CONRATH D W. Semantic similarity based on corpus statistics and lexical taxonomy[C]//Proc of International Conference on Research in Computational Linguistics Manchester. 1997:19-23.

(上接第 1681 页)

实验表明,不论在静态系统还是在动态系统中,HPHVF 与 HPEDF 的性能都分别优于其对应的传统算法。



(a)动态系统中各算法MDP分析　(b)动态系统中各算法AU分析

(c)动态系统中各算法VPP分析　(d)动态系统中各算法TV分析

图2　动态系统中四种算法的性能比较

5　结束语

系统过载加剧了资源竞争的程度,增大了作业级联抢占出现的概率,从而增大了任务的截止期错失率,降低了系统性能。有选择地执行作业,可以合理限制系统负载、减少无效抢占、降低资源竞争程度、提高资源利用率和系统收益。HP-OMS 通过确定作业的类型,拒绝执行那些无法完成的作业,使得系统性能得到最优化。HP-OMS 的主要创新点在于:解决了过载系统中作业级联抢占的现象;该策略能够适应调度任务集的动态改变。

参考文献:

[1] 黄文广,于士齐. 周期多帧任务的固定优先级调度算法的调度分析[J]. 计算机研究与发展,2001,38(2):240-245.

[2] NAGY S, BESTAVROS A. Value-cognizant admission control for RT-DB systems[C]//Proc of the 16th IEEE Real-time System Symposium. Washington DC: IEEE Computer Society,1996:230-239.

[3] ATLAS A, BESTAVROS A. Slack stealing job admission control [R]. Boston, MA:Boston University,1998.

[4] 夏家莉. 适用于嵌入式实时数据库系统的接纳控制机制 IACM [J]. 计算机学报,2004,27(3):295-301.

[5] De NIZ D, LAKSHMANAN K, RAJKUMAR R. On the scheduling of mixed-criticality real-time task sets[C]//Proc of the 30th IEEE Real-time Systems Symposium. Washington DC: IEEE Computer Society, 2009:291-300.

[6] 刘云生,李国徽. 基于预分析的实时任务处理[J]. 计算机研究与发展,2002,39(12):1728-1734.

[7] CHENG A M K, FENG Chen. Predictive thermal management for hard real-time tasks[J]. ACM SIGBED Review,2006,3(1):35-40.

[8] SCHRANZHOFER A,CHEN Jian-jia,THIELE L. Timing analysis for TDMA arbitration in resource sharing systems[C]//Proc of the 16th IEEE Real-time and Embedded Technology and Applications Symposium. Stockholm: IEEE Computer Society,2010:215-224.